M.Sc. in Data Science - Probability and Statistics for Data Analysis - Assignment 1

Spiros Politis

Nov. 2018

1. Assume that A and B are events of the sample space S for which we know:

$$2P(A) - P(A') = \frac{3}{5}, P(B|A) = \frac{5}{8} \text{ and } P(A|B) = \frac{4}{9}$$

Calculate the following probabilities:

(a) $P(A)$

(b) $P(A \cap B)$

(c) $P(B)$

(d) $P(A \cup B)$

(e) Are the events A and B independent?

Answers

a. $2P(A) - P(A') = \frac{3}{5} \implies 2P(A) - (1 - P(A)) = \frac{3}{5} \implies 3P(A) = \frac{8}{5} \implies P(A) = \frac{8}{15}$

b. $P(A \mid B) = \frac{P(A \cap B)}{P(B)} \implies P(A \cap B) = P(A|B)P(B)$ (1)

   Using Bayes rule to find $P(B)$ we have:

   $P(A|B) = P(B|A)\frac{P(A)}{P(B)} \implies P(B) = \frac{P(B|A)P(A)}{P(A|B)}$ (2)

   Substituting (2) into (1) we get:

   $P(A \cap B) = P(A|B)\frac{P(B|A)P(A)}{P(A|B)} \implies P(A \cap B) = P(B|A)P(A) \implies P(A \cap B) = \frac{5}{8}\frac{8}{15} \implies P(A \cap B) = \frac{1}{3}$

c. Applying Bayes rule we get:

   $P(A \mid B) = P(B \mid A)\frac{P(A)}{P(B)} \implies P(A \mid B)P(B) = P(B \mid A)P(A) \implies P(B) = \frac{P(B|A)P(A)}{P(A|B)} \implies P(B) = \frac{3}{4}$

d. $P(A \cup B) = P(A) + P(B) - P(A \cap B) \implies P(A \cup B) = \frac{8}{15} + \frac{3}{4} - \frac{1}{3} \implies P(A \cup B) = \frac{57}{60}$

e. The events are **not** independent, since $P(A \cap B) \neq \emptyset$

---

2. Two players, A and B, alternatively and independently flip a coin and the 1st player to obtain a head wins. Assume player A flips first.

(a) If the coin is fair, what is the probability that player A wins?

(b) More generally assume that $P(head) = p$ (not necessarily $\frac{1}{2}$). What is the probability that player A wins?

(c) Show that $\forall p$ such that $0 < p < 1$, we have that $P(\text{A wins}) > \frac{1}{2}$.

Answers

We have a discrete random variable X which is the oucome of flipping the coin.

a. The sequence of events for player A to win are the following:

$HT, TTH, TTTTH, TTTTTTH, \ldots$

which, define our sample space S.

Let $P(A_i)$ the probability that A produces H at the i-th toss, where i is odd:

$P(A_i) = (\frac{1}{2})^{i-1} \cdot \frac{1}{2} = (\frac{1}{2})^i$

Then, the total probability of A winning is:

$\sum_i P(A_i) = \sum_{\text{i odd}}^{\infty} (\frac{1}{2})^i = \sum_{i=0}^{\infty} (\frac{1}{2})^{2i+1} = \frac{1}{2} \sum_{i=0}^{\infty} (\frac{1}{4})^i = \frac{1}{2} \cdot \frac{4}{3} = \frac{2}{3}$

b. In the general case, where $P(head) = p$, the probability of A winning at the i-th toss would be:

$P(A_i) = p^{i-1} \cdot p = p^i$

Then, the total probability of A winning is:

$\sum_i P(A_i) = \sum_{\text{i odd}}^{\infty} p^i = \sum_{i=0}^{\infty} p^{2i+1} = p \sum_{i=0}^{\infty} p^i = p \cdot \frac{4}{3}$

c.

---

3. A telegraph signals “dot” and “dash” sent in the proportion 3:4, where erratic transmission cause a dot to become dash with probability 1/4 and a dash to become a dot with probability 1/3.

(a) If a dash is received, what is the probability that a dash has been sent?

(b) Assuming independence between signals, if the message dot-dot was received, what is the probability distribution of the four possible messages that could have been sent?

Answers

Let us define the following events:

dot_s: a dot was sent by the transmitter
$dash_s$: a dash was sent by the transmitter
dot_r: a dot was received
$dash_r$: a dash was received

Given the dot and dash ratio of $\frac{3}{4}$, we will use this function to deduce the probabilitites of sending each symbol. It holds that

$$\frac{P(dot_s)}{P(dash_s)} = \frac{3}{4} \implies 4P(dot_s) = 3P(dash_s) \implies 4P(dot_s) = 3(1 - P(dot_s)) \implies P(dot_s) = \frac{3}{7}$$

$$P(dash_s) = 1 - P(dot_s) \implies P(dash_s) = 1 - \frac{3}{7} \implies P(dash_s) = \frac{4}{7}$$

a. Let us utilize Bayes rules to compute the probability of receiving a dash given a dash was sent:

$$P(dash_s|dash_r) = P(dash_r|dash_s)\frac{P(dash_s)}{P(dash_r)} = \frac{P(dash_r|P(dash_s))P(dash_s)}{P(dash_r|dash_s)P(dash_s) + P(dash_r|dot_s)P(dot_s)} = \frac{\frac{2}{3}\frac{4}{7}}{\frac{2}{3}\frac{4}{7} + \frac{1}{4}\frac{3}{7}} = \frac{32}{41} \approx 0.78$$

b. We identify the following probabilities of message combinations:

$P(dot_s|dot_r)P(dot_s|dot_r)$ (1)
$P(dot_s|dot_r)P(dash_s|dot_r)$ (2)
$P(dash_s|dot_r)P(dot_s|dot_r)$ (3)
$P(dash_s|dot_r)P(dash_s|dot_r)$ (4)

The uniquely identifiable probabilities, for which we need to apply Bayes rules and substitute above, are:

$$P(dot_s|dot_r) = \frac{P(dot_r|dot_s)P(dot_s)}{P(dot_r)} = \frac{P(dot_r|dot_s)P(dot_s)}{P(dot_r|dot_s)P(dot_s)+P(dot_r|dash_s)P(dash_s)} = \frac{\frac{3}{4}\frac{3}{7}}{\frac{3}{4}\frac{3}{7}+\frac{1}{3}\frac{4}{7}} = \frac{5292}{8428} \approx 0.628 \quad (5)$$

$$P(dash_s|dot_r) = \frac{P(dot_r|dash_s)P(dash_s)}{P(dot_r)} = \frac{P(dot_r|dash_s)P(dash_s)}{P(dot_r|dot_s)P(dot_s)+P(dot_r|dash_s)P(dash_s)} = \frac{\frac{1}{3}\frac{4}{7}}{\frac{3}{4}\frac{3}{7}+\frac{1}{3}\frac{4}{7}} = \frac{2352}{8428} \approx 0.372 \quad (6)$$

Substituting (1) with (5), (5) we get: $P(dot_s|dot_r)P(dot_s|dot_r) = 0.628 \cdot 0.628 \approx 0.394$
Substituting (2) with (5), (6) we get: $P(dot_s|dot_r)P(dash_s|dot_r) = 0.628 \cdot 0.372 \approx 0.234$
Substituting (3) with (6), (5) we get: $P(dash_s|dot_r)P(dot_s|dot_r) = 0.372 \cdot 0.628 \approx 0.234$
Substituting (4) with (6), (6) we get: $P(dash_s|dot_r)P(dash_s|dot_r) = 0.372 \cdot 0.372 \approx 0.138$

Sanity check:
$P(dot_s|dot_r)P(dot_s|dot_r) + P(dot_s|dot_r)P(dash_s|dot_r) + P(dash_s|dot_r)P(dot_s|dot_r) + P(dash_s|dot_r)P(dash_s|dot_r) = 1$

The probabilities sum up to 1, therefore we have a probability distribution of receiving a dot-dot message.

---

4. Let X be a continuous random variable with pdf $f(x)$ and cdf $F(x)$. For a fixed number x_0 (such that $F(x_0) < 1$), define the function:

$$g(x) = \begin{cases} \frac{f(x)}{1-F(x_0)} & x \geq x_0 \\ 0 & x < x_0 \end{cases}$$

Prove that $g(x)$ is a pdf (also known as hazard function).

Answers

For $g(x)$ to be a PDF, the following two conditions must apply:

$$g(x) \geq 0, \forall x \text{ (1)}$$

and

$$\int_{-\infty}^{+\infty} g(x)\ dx = 1 \text{ (2)}$$

We also know that:

$$f(x) = \frac{\partial(F(x))}{\partial x}$$

and

$$F(x) = \int_{-\infty}^{+\infty} f(x)$$

since $F(x)$ is the CDF of $f(x)$, a relationship that will help us with our calculations.

So, for (1) we have $F(x_0) < 1 \implies F(x_0) - 1 < 0 \implies 1 - F(x_0) > 0$ and $f(x) \geq 0$ because $f(x)$ is a PDF. Therefore, the quantity $\frac{f(x)}{1-F(x_0)} \geq 0 \implies g(x) \geq 0$.

For (2) we have:

$\int_{-\infty}^{+\infty} g(x)\ dx = 1 \implies$
$\int_{0}^{x_0} 0\ dx + \int_{x_0}^{+\infty} \frac{f(x)}{1-F(x_0)}\ dx = 1 \implies$
$0 + \frac{1}{1-F(x_0)} \cdot \int_{x_0}^{+\infty} f(x)\ dx = 1 \implies$
$\frac{F(x)}{1-F(x)}\Big|_{x_0}^{+\infty} = 1$

---

5. Consider a telephone operator who, on the average, handles five calls every three minutes.
(a) What is the probability of no calls in the next minute?
(b) What is the probability of at least two calls in the next minute?
(c) What is the probability of at most two calls in the next five minutes?

Answers

Number of calls is a discrete random variable distributed as $X \sim Pois(x|\lambda)$. The distribution PMF is $f(x \mid \lambda) = \frac{e^{-\lambda}\lambda^x}{x!}$, with the parameter λ being $\lambda = \frac{5}{3}$.

a. We are looking for $P(X=0) = \frac{e^{-\frac{5}{3}} \frac{5}{3}^0}{0!} \approx 0.189$

b. We are looking for $P(X \geq 2) = 1 - (P(X=0) + P(x=1)) = 1 - (\frac{e^{-\frac{5}{3}} \frac{5}{3}^0}{0!} + \frac{e^{-\frac{5}{3}} \frac{5}{3}^1}{1!} = 1 - (0.189 + 0.315) \approx 0.496$

c. The parameter λ we will use is: $\lambda = \frac{5}{3} * 5 \implies \lambda = \frac{25}{3}$

We are looking for the probability

$$P(X=0) + P(X=1) + P(X=2) = \frac{e^{-\frac{25}{3}} \frac{25}{3}^0}{0!} + \frac{e^{-\frac{25}{3}} \frac{25}{3}^1}{1!} + \frac{e^{-\frac{25}{3}} \frac{25}{3}^2}{2!} = 0.00024 + 0.002 + 0.008 \approx 0.01024$$

---

6. Let $X_1, X_2, \ldots, X_n$ be a random sample form a $Gamma(\alpha, \beta)$ distribution. Find a two-dimensional sufficient statistic for (α, β).

Answers

The PDF of the Gamma distribution is:

$$f(x|\alpha, \beta) = \frac{1}{\Gamma(\alpha) \cdot \beta^{\alpha}} \cdot x^{(\alpha-1)} \cdot e^{-\frac{x}{\beta}}$$

A minimum sufficient statistic of (α, β) would be a function $T(x)$ iff $\frac{f_\theta(x)}{f_\theta(y)}$ is independent of θ.

Therefore:

$$\frac{f_\theta(x_n)}{f_\theta(y_n)} = \frac{\prod_{i=1}^{n} \frac{1}{\Gamma(\alpha) \cdot \beta^{\alpha}} \cdot x_n^{(\alpha-1)} \cdot e^{-\frac{xn}{\beta}}}{\prod_{i=1}^{n} \frac{1}{\Gamma(\alpha) \cdot \beta^{\alpha}} \cdot y_n^{(\alpha-1)} \cdot e^{-\frac{yn}{\beta}}} =$$
$$\frac{\left(\frac{1}{\Gamma(\alpha) \cdot \beta^{\alpha}} \cdot x_1^{\alpha-1} \cdot e^{-\frac{x1}{\beta}}\right) \cdot \left(\frac{1}{\Gamma(\alpha) \cdot \beta^{\alpha}} \cdot x_2^{\alpha-1} \cdot e^{-\frac{x2}{\beta}}\right) \cdot \ldots \cdot \left(\frac{1}{\Gamma(\alpha) \cdot \beta^{\alpha}} \cdot x_n^{\alpha-1} \cdot e^{-\frac{xn}{\beta}}\right)}{\left(\frac{1}{\Gamma(\alpha) \cdot \beta^{\alpha}} \cdot y_1^{\alpha-1} \cdot e^{-\frac{y1}{\beta}}\right) \cdot \left(\frac{1}{\Gamma(\alpha) \cdot \beta^{\alpha}} \cdot y_2^{\alpha-1} \cdot e^{-\frac{y2}{\beta}}\right) \cdot \ldots \cdot \left(\frac{1}{\Gamma(\alpha) \cdot \beta^{\alpha}} \cdot y_n^{\alpha-1} \cdot e^{-\frac{yn}{\beta}}\right)} =$$
$$\frac{\left(\frac{1}{\Gamma(\alpha) \cdot \beta^{\alpha}}\right) \cdot (\prod_{i=1}^{n} x_i)^{\alpha-1} \cdot e^{-\frac{\sum_{i=1}^{n} x_i}{\beta}}}{\left(\frac{1}{\Gamma(\alpha) \cdot \beta^{\alpha}}\right) \cdot (\prod_{i=1}^{n} y_i)^{\alpha-1} \cdot e^{-\frac{\sum_{i=1}^{n} y_i}{\beta}}} =$$
$$\left(\frac{\prod_{i=1}^{n} x_i}{\prod_{i=1}^{n} y_i}\right)^{\alpha-1} \cdot e^{-\frac{\sum_{i=1}^{n} x_i + \sum_{i=1}^{n} x_i}{\beta}}$$

We have therefore identified that $T(x)$ is constant with respect to α when the products are the same and constant with respect to β when the sums are the same. So, our minimum sufficient statistic for (α, β) is $(\prod_{i=1}^{n} x_i, \sum_{i=1}^{n} x_i)$.

---

7. One observation X is taken from a $N(0, \sigma^2)$ distribution.

(a) Find an unbiased estimate of σ^2.

(b) Find the maximum likelihood estimator (MLE) of σ^2.

Answers

a. A well-known estimator for σ^2 is $s^2 = \frac{1}{n-1} \sum_{i=1}^{n} (x_i - \overline{x})^2$.

We know that:

$E[X] = \mu$ (1)

$\sigma^2 = E[X^2] - (E[X])^2 \implies \sigma^2 = E[X^2] - \mu^2 \implies E[X^2] = \sigma^2 + \mu^2$ (2)

We need to show that s^2 is an unbiased estimator for σ^2, therefore $E[s^2] = \sigma^2$:

$E[s^2] =$
$E[\frac{1}{n}\sum_{i=1}^{n}(x_i-\overline{x})^2] =$
$\frac{1}{n-1}E[\sum_{i=1}^{n}(x_i-\overline{x})(x_i-\overline{x})] =$
$\frac{1}{n-1}E[\sum_{i=1}^{n}x_i^2-2x_i\overline{x}+\overline{x}^2] =$
$\frac{1}{n-1}E[\sum_{i=1}^{n}x_i^2-\sum_{i=1}^{n}2x_i\overline{x}+\sum_{i=1}^{n}\overline{x}^2] =$
$\frac{1}{n-1}E[\sum_{i=1}^{n}x_i^2-2\overline{x}\sum_{i=1}^{n}x_i+n\overline{x}^2] =$
$\frac{1}{n-1}E[\sum_{i=1}^{n}x_i^2-2n\overline{x}^2+n\overline{x}^2] =$
$\frac{1}{n-1}E[\sum_{i=1}^{n}x_i^2-n\overline{x}^2] =$
$\frac{1}{n-1}\left[\sum_{i=1}^{n}E[x_i^2]-E[n\overline{x}^2]\right] =$
(Substituting from (2))
$\frac{1}{n-1}\left[\sum_{i=1}^{n}(\sigma^2+\mu^2)-nE[\overline{x}^2]\right] =$
$\frac{1}{n-1}\left[n\sigma^2+n\mu^2-n(\frac{\sigma^2}{n}+\mu^2)\right] =$
$\frac{1}{n-1}(n\sigma^2+n\mu^2-\sigma^2-n\mu^2) =$
$\frac{1}{n-1}(n\sigma^2-\sigma^2) =$
$\frac{\sigma^2(n-1)}{n-1} = \sigma^2$

Therefore, s^2 is indeed an unbiased estimator of σ^2.

b. Let us first note the PDF for the Normal distribution, which is: $f(x|\mu,\sigma^2) = \frac{1}{\sqrt{2\pi\sigma^2}}e^{-\frac{(x-\mu)^2}{2\sigma^2}}$

The estimators of the PDF are then symbolized as $\theta_1=\mu$, $\theta_2=\sigma^2$, therefore we should rewrite the PDF for the Normal distribution as:
$f(x|\theta_1,\theta_2) = \frac{1}{\sqrt{2\pi\theta_2}}e^{-\frac{(x-\theta_1)^2}{2\theta_2}}$

Let us write down the likelihood function:

$L(\theta_1,\theta_2|\underline{x}) = \prod_{i=1}^{n}\frac{1}{\sqrt{2\pi\theta_2}}\cdot e^{-\frac{(x_i-\theta_1)^2}{2\theta_2}} =$
$\frac{1}{\sqrt{2\pi\theta_2}}\cdot e^{-\frac{(x_1-\theta_1)^2}{2\theta_2}}\cdot\frac{1}{\sqrt{2\pi\theta_2}}\cdot e^{-\frac{(x_2-\theta_1)^2}{2\theta_2}}\cdot\ldots\cdot\frac{1}{\sqrt{2\pi\theta_2}}\cdot e^{-\frac{(x_n-\theta_1)^2}{2\theta_2}} =$
$(\frac{1}{\sqrt{2\pi\theta_2}})^n\cdot e^{-\sum_{i=1}^{n}\frac{(x_i-\theta_1)^2}{2\theta_2}} =$
$(\frac{1}{2\pi\theta_2})^{\frac{n}{2}}\cdot e^{-\sum_{i=1}^{n}\frac{(x_i-\theta_1)^2}{2\theta_2}}$

Let us take the log of the likelihood function:

$l(L(\theta_1,\theta_2|\underline{x})) = \sum_{i=1}^{n}ln(\frac{1}{\sqrt{2\pi\theta_2}}\cdot e^{-\frac{(x_i-\theta_1)^2}{2\theta_2}}) =$
$\sum_{i=1}^{n}\left[ln(\frac{1}{\sqrt{2\pi\theta_2}})+ln(e^{-\frac{(x_i-\theta_1)^2}{2\theta_2}})\right] =$
$\sum_{i=1}^{n}\left[ln((2\pi\theta_2)^{-\frac{1}{2}})-\frac{(x_i-\theta_1)^2}{2\theta_2}ln(e)\right] =$
$\sum_{i=1}^{n}\left[-\frac{1}{2}ln(2\pi\theta_2)-\frac{(x_i-\theta_1)^2}{2\theta_2}\right] =$
$-\frac{n}{2}ln(2\pi\theta_2)-\frac{1}{2\theta_2}\sum_{i=1}^{n}(x_i-\theta_1)^2 =$
$-\frac{n}{2}ln(2\pi)-\frac{n}{2}ln(\theta_2)-\frac{\sum_{i=1}^{n}(x_i-\theta_1)^2}{2\theta_2}$

Taking the derivative of the log likelihood function with respect to θ_2 we get:

$\frac{\partial l}{\partial\theta_2} = (-\frac{n}{2}ln(2\pi)-\frac{n}{2}ln(\theta_2)-\frac{\sum_{i=1}^{n}(x_i-\theta_1)^2}{2\theta_2})' =$
$0-\frac{n}{2\theta_2}+\frac{\sum_{i=1}^{n}(x_i-\theta_1)^2}{2\theta_2^2} =$
$-\frac{n}{2\theta_2}+\frac{\sum_{i=1}^{n}(x_i-\theta_1)^2}{2\theta_2^2}$

Solving for $\frac{\partial l}{\partial \theta_2} = 0$ we get:

$$
\begin{aligned}
&-\frac{n}{2\theta_2} + \frac{\sum_{i=1}^{n}(x_i-\theta_1)^2}{2\theta_2^2} = 0 \\
&-\frac{1}{2\theta_2}(n - \frac{\sum_{i=1}^{n}(x_i-\theta_1)^2}{\theta_2}) = 0 \\
&n = \frac{\sum_{i=1}^{n}(x_i-\theta_1)^2}{\theta_2} \implies \\
&\hat{\theta}_2 = \frac{\sum_{i=1}^{n}(x_i-\theta_1)^2}{n}
\end{aligned}
$$

We have shown that the maximum likelihood estimator of σ^2 for the Normal probability distribution is $\hat{\theta}_2 = \frac{\sum_{i=1}^{n}(x_i-\theta_1)^2}{n}$.

We also need to verify that it is correct by taking the second partial derivative, with respect to θ_2, of the likelihood function and making sure it is negative:

$$
\begin{aligned}
&\frac{\partial^2 l}{\partial^2 \theta_2} = (-\frac{n}{2\theta_2} + \frac{\sum_{i=1}^{n}(x_i-\theta_1)^2}{2\theta_2^2})' == \\
&-\frac{n}{2}(\theta_2^{-1})' + \frac{\sum_{i=1}^{n}(x_i-\theta_1)^2}{2}(\theta_2^{-2})' = \\
&\frac{n}{2\theta_2^2} - \frac{\sum_{i=1}^{n}(x_i-\theta_1)^2}{\theta_2^3}
\end{aligned}
$$

Therefore:

$$
\begin{aligned}
&\frac{n}{2\theta_2^2} - \frac{\sum_{i=1}^{n}(x_i-\theta_1)^2}{\theta_2^3} < 0 \implies \\
&\frac{n}{2} < \frac{\sum_{i=1}^{n}(x_i-\theta_1)^2}{\theta_2} \implies \\
&\frac{n}{2} < \frac{\sum_{i=1}^{n}(x_i-\theta_1)^2}{\theta_2} \implies \\
&\theta_2 < \frac{2 \cdot \sum_{i=1}^{n}(x_i-\theta_1)^2}{n} \implies \\
&\theta_2 < 2 \cdot \theta_2 \text{ (because } \theta_2 = \frac{\sum_{i=1}^{n}(x_i-\theta_1)^2}{n})
\end{aligned}
$$

which is true.

As a final note, because it is given that $X \sim N(0, \sigma^2)$, therefore $\mu = 0$, the maximum likelihood estimator $\hat{\theta}_2$ for our particular case becomes

$$\hat{\theta}_2 = \frac{\sum_{i=1}^{n} x_i^2}{n}$$

---

8. Two random samples of size of n = 10 from a process producing bottles of water are gathered. The sample means are $\overline{x_1}$ = 1000.42ml and $\overline{x_2}$ = 999.58ml respectively. We assume that the data are normally distributed with σ = 0.62 (known).

(a) Provide a confidence interval for the mean of each subgroup in α = 0.05 significance level.

(b) Test if the sample means of the subgroups are statistically equal in α = 0.05 significance level.

(c) Test if $\overline{x_1}$ is statistically greater than 1Litre in α = 0.05 significance level.

Answers

a. We are interested in constructing a CI for each of the unknown population parameters of each subgroup, namely μ_1 and μ_2. Let us symbolize the CIs as CI_1 and CI_2 respectively. Then we will have:

$$CI_1 = \overline{x_1} \pm ME_1 \quad (1)$$

$$CI_2 = \overline{x_2} \pm ME_2 \quad (2)$$

Since the significance level $\alpha = 0,05$, we are looking for a CI at the confidence level of $1 - \alpha = 1 - 0,05 = 0,95$. Also, we need to lookup the Z value of the standard normal distribution, using perhaps a Z-table or software, at which the probability of Z is $P(\frac{a}{2} \leq Z \leq 1 - \frac{a}{2})$. This value is 1.96.

Therefore, for (1) we have:

$$CI_1 = \overline{x_1} \pm \left(Z_{\frac{\alpha}{2}} \cdot \frac{\sigma}{\sqrt{n}}\right) \implies$$
$$CI_1 = 1000.42 \pm \left(1.96 \cdot \frac{0.62}{\sqrt{10}}\right) \implies$$
$$CI_1 = 1000.42 \pm (1.96 \cdot 0.196) \implies$$
$$CI_1 = (1000.036, 1000.804)$$

Interpreting our results, we can state that 95% of the time, random sampling obtained from the water bottle producing process will yield the true population mean μ, which will lie in the $(1000.036ml, 1000.804ml)$ interval.

For (2) we have:

$$CI_2 = \overline{x_2} \pm \left(Z_{\frac{\alpha}{2}} \cdot \frac{\sigma}{\sqrt{n}}\right) \implies$$
$$CI_2 = 999.58 \pm \left(1.96 \cdot \frac{0.62}{\sqrt{10}}\right) \implies$$
$$CI_2 = 999.58 \pm (1.96 \cdot 0.196) \implies$$
$$CI_2 = (999.196, 999.964)$$

Interpreting our results, we can state that 95% of the time, random sampling obtained from the water bottle producing process will yield the true population mean μ, which will lie in the $(999.196ml, 999.964ml)$ interval.

b. We have a case of hypothesis testing for independent samples and our hypothesis testing framework is as follows:

$H_0 : \mu_1 = \mu_2$: sample means of the subgroups are statistically equal
$H_A : \mu_1 \neq \mu_2$: sample means of the subgroups are **not** statistically equal

The applicable formula for finding the Z test statistic is:

$$Z = \frac{(\overline{x_1} - \overline{x_2}) - (\mu_1 - \mu_2)}{\sqrt{\frac{\sigma_1^2}{n_1} + \frac{\sigma_1^2}{n_2}}} = \frac{(\overline{x_1} - \overline{x_2}) - 0}{\sqrt{\frac{\sigma_1^2}{n_1} + \frac{\sigma_1^2}{n_2}}}$$

Therefore:

$$Z = \frac{1000.42 - 999.58}{\sqrt{\frac{(0.62)^2}{10+10}}} = 3.03$$

The value of our Z test statistic is larger than $Z_{\frac{1-\alpha}{2}} = 1.96$ (looked up from a Z-table). We therefore conclude that we should **reject** H_0 (that the sample means of the subgroups are statistically equal) in favor of H_A at significance level $\alpha = 0.05$.

c. We need to perform a test of significance on our data. We begin by stating our hypotheses as:

$H_0 : \mu = 1\text{Litre}$
$H_0 : \mu > 1\text{Litre}$

Our test will provide us with the probability of observed or more extreme outcome under H_0, given our data.

We proceed by computing our test statistic and finding the $p - value$:

$$z = \frac{\overline{x_1} - \mu}{\frac{\sigma}{\sqrt{n}}} \implies$$
$$z = \frac{1000.42 - 1000}{\frac{0.62}{\sqrt{10}}} \implies$$
$$z = 2.142$$

Since this is a one-sided test, the $P - value$ is equal to the probability of observing a value greater than 2.142 in the standard normal distribution, or $P(Z > 2.142) = 1 - P(Z < 2.142) = 1 - 0.9838 = 0.0162$.

The $P - value$ is less than $\alpha = 0.05$, indicating that it is highly unlikely that these results would be observed under the null hypothesis. We therefore reject H_0 in favor of H_A and conclude that our sample mean $\overline{x_1}$ is statistically greater than $1Litre$ at the $\alpha = 0.05$ significance level.